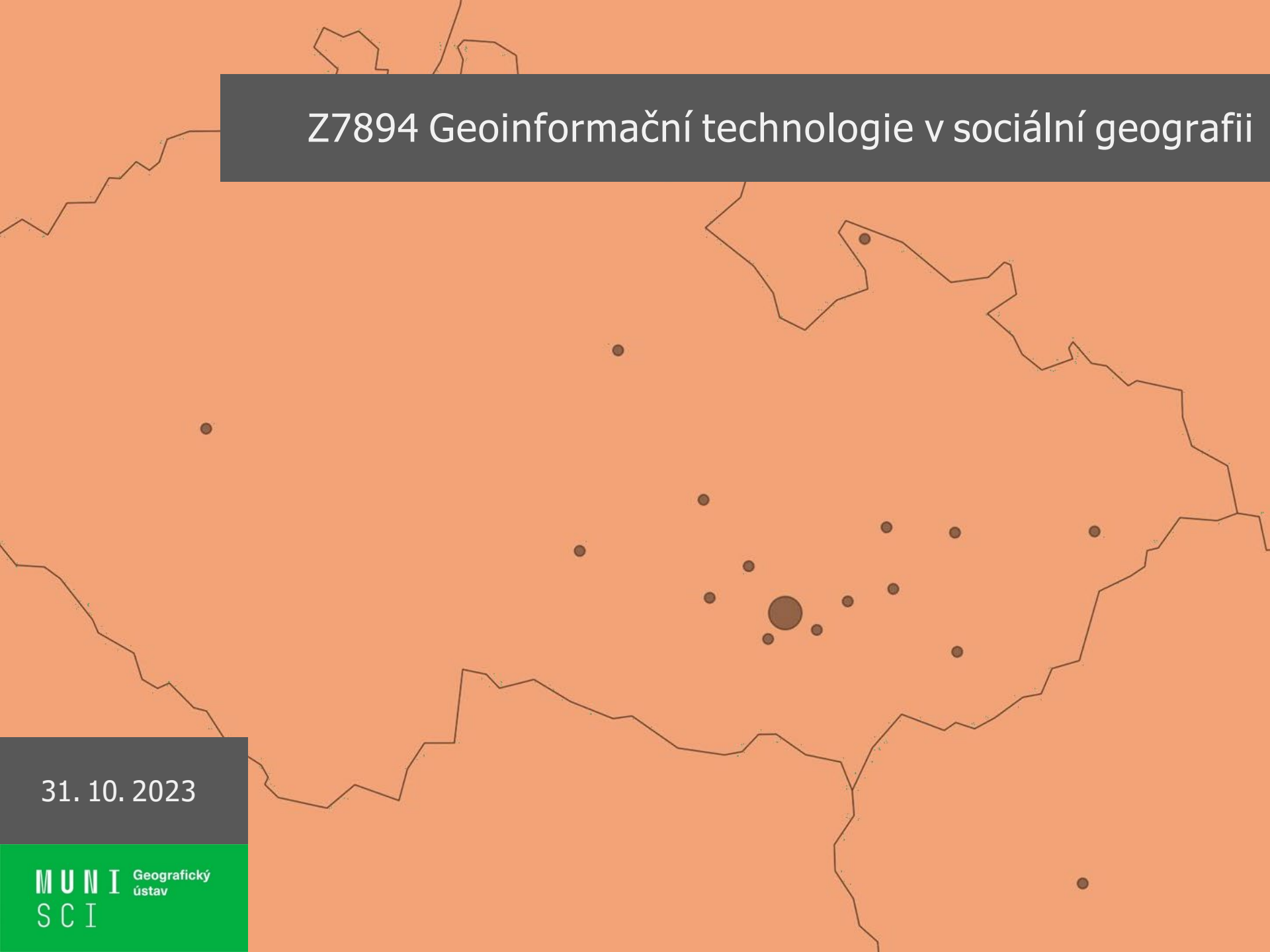
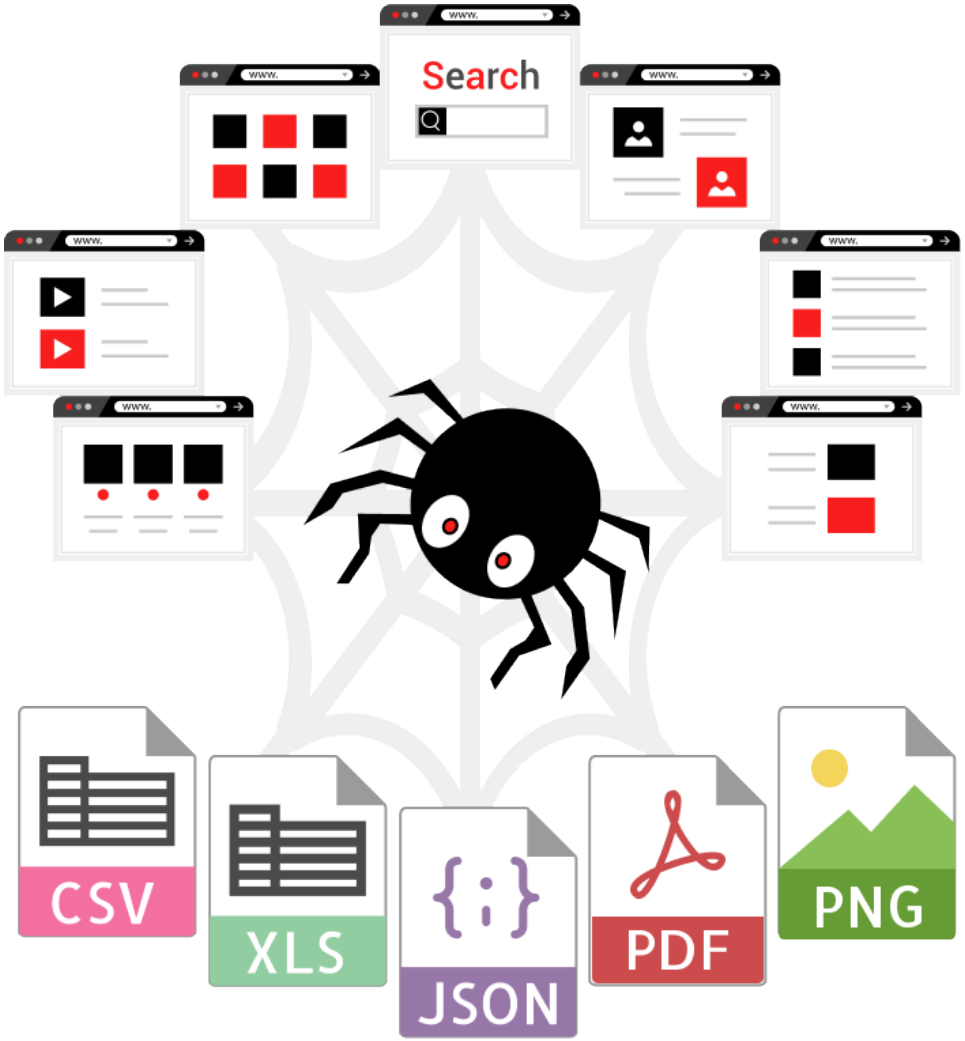
Z7894 Geoinformační technologie v sociální geografii



31. 10. 2023

MUNI SCI Geografický ústav

Web scraping: teorie

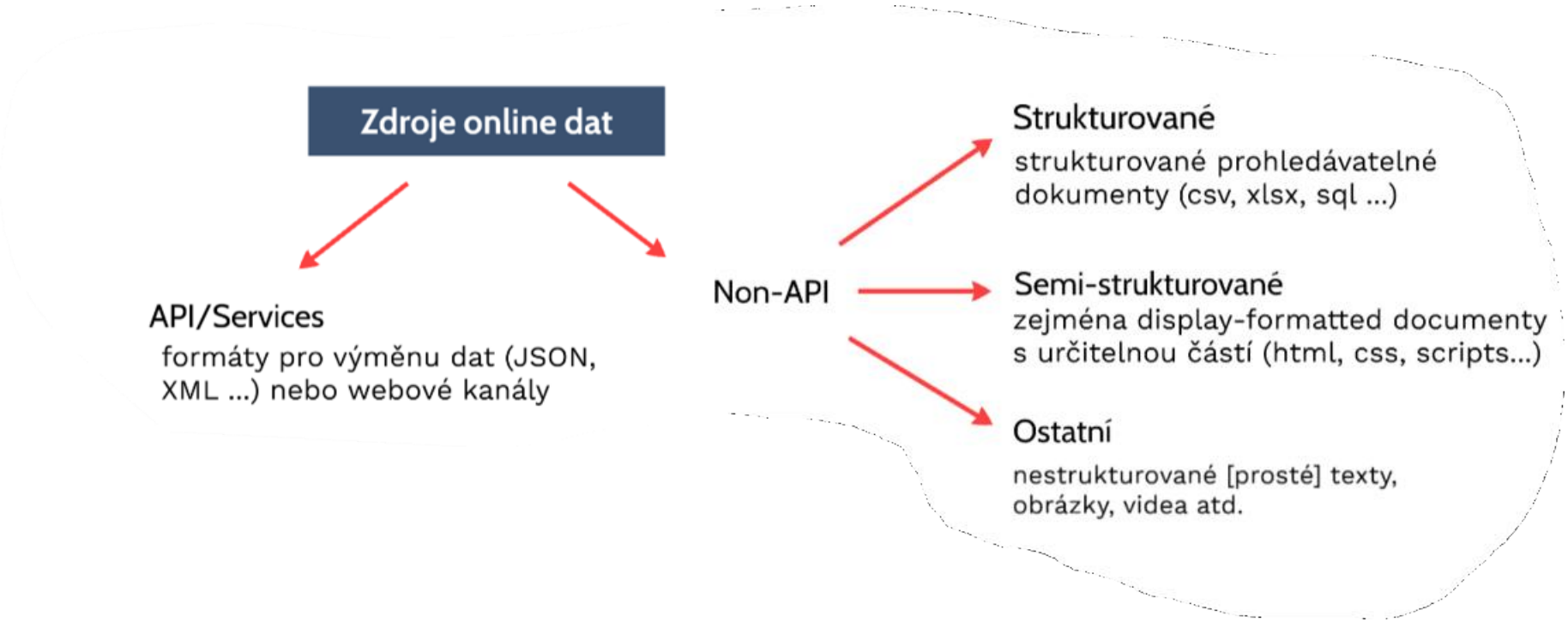


Web scraping

- **Získávání dat** z webových stránek (ale i webových aplikací a jejich součástí) a jejich **ukládání do strojově zpracovatelné podoby** (např. tabulky, samostatná databáze apod.).
- Související pojmy: **Extrakce dat** a **automatizace**
- Sporná je otázka (i)legality web scrapingu – někdy jde **o šedou zónu**

- Možné usecases pro geografy:
  - kopírování obsáhlých tabulek s listingem (např. data ČSÚ, ERU…),
  - ukládání webových seznamů do tabulky (např. seznam chystaných akcí v Brně…),
  - automatické načítání údajů (např. „excelovský sheet" napojený aktuální teplotu),
  - periodické ukládání určitého údaje (např. návštěvnost bazénu),
  - …

Typy zdrojů online dat na webu ke scrapování



Web scraping se za nejčastěji považuje odchytávání elementů html (resp. **parsování html**), ale může jít i o další formy spojené s **reverzním inženýrstvím** aplikací a extrakcí dat.

Typy (časoprostorových) geografických dat

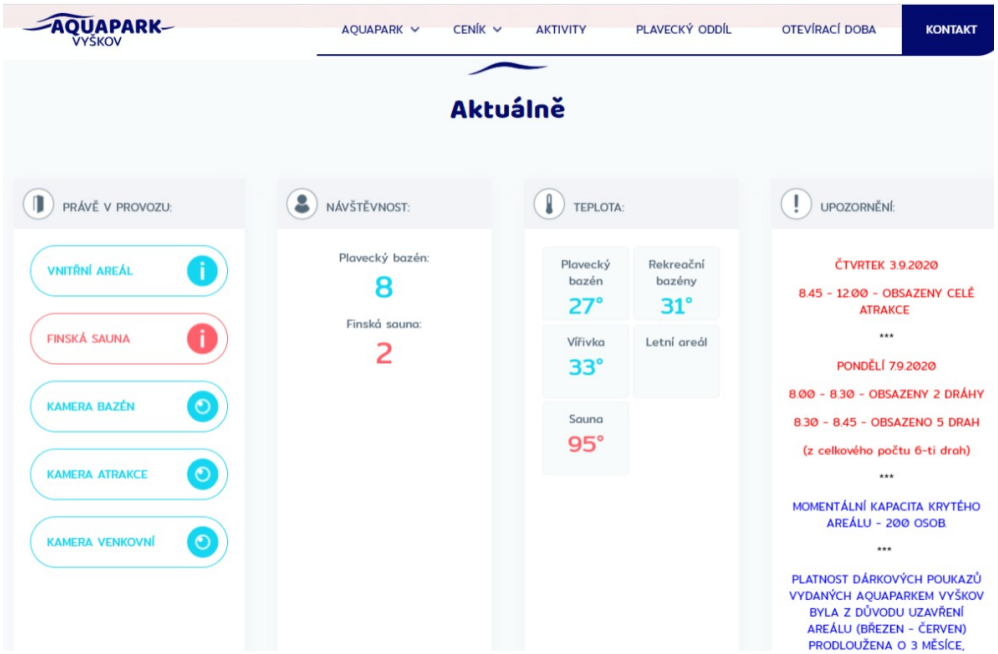
již zaznamenaná měření teploty
vs
aktuální teplota
vs
predikce budoucí teploty

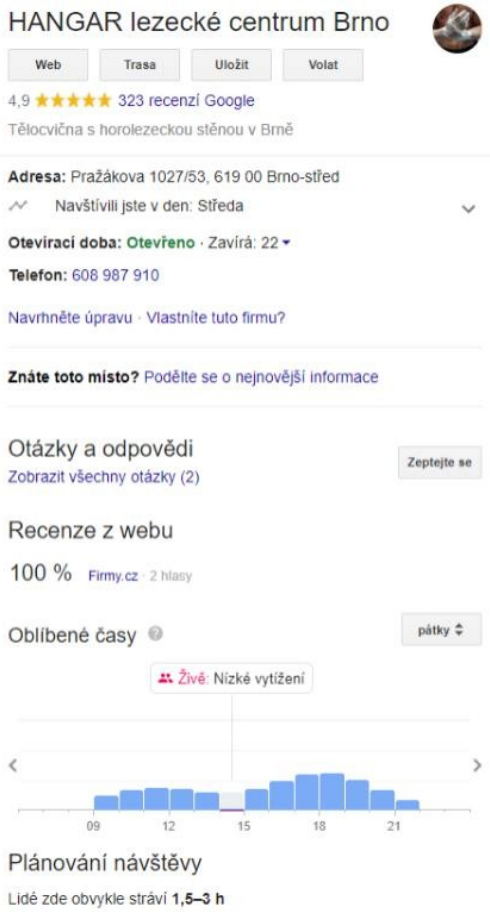
aktuální obsazenost parkoviště
vs
google popular times

návštěvnost v procentech
vs
návštěvnost přesná

Některé scrapování je nutné **automatizovat** (např. pokud chceme sbírat údaj o návštěvnosti v čase a sestavit časovou řadu).

Typy časoprostorových dat - ukázka





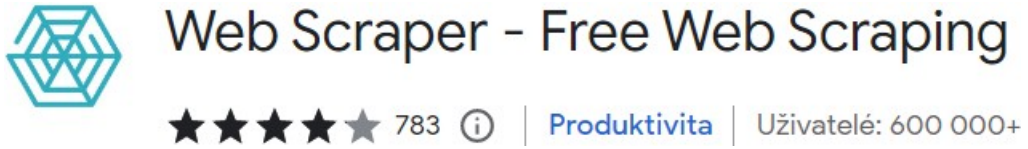
Web scraping: Nástroje

Základní rozdělení

- **Jednoduché s GUI** – např. Web Scraper plugin pro webový prohlížeč
- **Složitější pro programovací jazyky** – např. knihovny pro python/R (např. rvest)
- **Vlastní skripty/programy**
- **Komerční cloudové platformy** – např. Apify, Octoparse, Scrapestack

Jaké jsou výhody a nevýhody jednotlivých řešení?

Zadání cvičení
Praktická ukázka



Web Scraper s GUI ve vývojářském režimu Chrome

https://chrome.google.com/webstore/detail/web-scraper-free-web-scra/jnhgnonknehpejjnehehllkliplmbmhn

Vytvoření schémat, resp. jejich import a zahájení scrapování.

Dotazy?

Děkuji za pozornost